Learning with the Neyman–Pearson and min–max criteria

Los Alamos
National Laboratory

Adam Cannon
Department of Computer Science
Columbia University
New York, NY 10027, USA
cannon@cs.columbia.edu

James Howse, Don Hush and Clint Scovel
Modeling, Algorithms and Informatics Group, CCS-3
Mail Stop B265
Los Alamos National Laboratory
Los Alamos, NM 87545
{jhowse,dhush,jcs}@lanl.gov

LANL Technical Report: LA–UR–02-2951

Report Date: June 21, 2002

Abstract

We study two design criteria for classification: the Neyman–Pearson criterion and a min–max criterion. For each we prove a lemma bounding estimation error in terms of error deviance. We then show how these lemmas can be used to determine probabilistic guarantees on estimation error.

1 Introduction

Consider a set X, a finite set Y, and a probability space $Z = X \times Y$ and let $z = (x, y)$ denote the corresponding random variable with probability measure $\mathcal{P}$, conditional probability measures $\mathcal{P}_y$ and y-marginal probability measure P. Although the results of this paper are quite general, for exposition purposes we consider the classification problem. In particular, let $\mathcal{F}$ denote a class of functions(classifiers) $f : X \to Y$. The generalization error of the classifier f is defined as

$$e(f) = \mathcal{P}(f(x) \neq y).$$

We now suppose that n samples are taken from Z. We refer to the vector $(z_1, z_2, .., z_n)$ of n samples as an n-sample in Z^n. The most common way of selecting n-samples is independent identically distributed (*i.i.d.*) sampling but we shall use others in this paper such as retrospective sampling. Consequently, to distinguish the measure used to describe the sample plan from the measure used to compute generalization error, we denote the probability measure on Z^n corresponding to the sample plan as $\mathcal{P}_S$. When the sample plan is *i.i.d* from $\mathcal{P}$, we define the empirical error in terms of an n-sample as the fraction of misclassified data samples

$$\hat{e}(f) = \frac{1}{n}\sum_{i=1}^{n} I\big(f(x_i) \neq y_i\big).$$

The most common goal in classification is to select a classifier from $\mathcal{F}$

$$f^* \in \arg\min_{f \in \mathcal{F}} e(f) \tag{1}$$

that minimizes the generalization error $e(f)$. Throughout this paper we ignore questions of whether minima or maxima are actually attained. This detail is easy to include by introducing approximation parameters and approximate minima/maxima but obscures the presentation. Suppose we employ the well–known design strategy that chooses a classifier

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \hat{e}(f) \tag{2}$$

that minimizes the empirical error $\hat{e}(f)$. The difference

$$e(\hat{f}) - e(f^*)$$

between the generalization error of the chosen classifier and the optimal generalization error for the family $\mathcal{F}$ is called the *estimation error* and the largest possible absolute difference

$$\sup_{f \in \mathcal{F}} |e(f) - \hat{e}(f)|$$

between the generalization error and the empirical error is called the *error deviance*. A lemma of (Vapnik & Chervonenkis, 1974) states that

$$e(\hat{f}) - e^* \leq 2 \sup_{f \in \mathcal{F}} |e(f) - \hat{e}(f)|.$$

That is, the estimation error of a classifier designed by minimizing the empirical error is bounded by twice the error deviance. This lemma allows results on the convergence of empirical processes to generate probabilistic guarantees on the estimation error, such as the celebrated Vapnik–Chervonenkis theorem. Because this simple lemma transforms results from empirical process theory, for which much is known, to results about a learning paradigm, we refer to this lemma as a *fundamental lemma*.

In this paper we provide *fundamental lemmas* for the Neyman–Pearson and min–max error criteria for classifiers built by minimizing empirical versions of these criteria. We then show how these lemmas lead to probabilistic guarantees on the estimation error for such classifiers. We ignore questions of measurability.

2 The Neyman–Pearson Problem

In this section we let $Y = \{0, 1\}$ and we define the class conditional errors e_i by

$$e_i(f) = \mathcal{P}_i(f(x) \neq i), \;\; i \in Y.$$

The Neyman–Pearson problem is motivated by real world scenarios where it is important that one of these errors be no greater than some fixed value. For example in fraud detection the classification system often has no utility unless the false alarm rate (i.e. the rate at which fraudulent activity is predicted when it is not present) can be kept below a fixed level. In the Neyman–Pearson problem we impose a this type of constraint on one of the errors and optimize the other. The version of the Neyman–Pearson problem we treat here is(see e.g. (Van Trees, 1968))

$$\begin{array}{ll} \min_{f \in \mathcal{F}_1} & e_0(f) \\ \text{where} & \mathcal{F}_1 = \{f : f \in \mathcal{F}, e_1(f) \leq \alpha\} \end{array} \tag{3}$$

for some $\alpha \in [0, 1]$. We would like to formulate an empirical optimization problem in such a way that we can bound the estimation error of its solution. To this end let $\hat{e}_0$ and $\hat{e}_1$ be the class conditional empirical errors defined by

$$\hat{e}_i(f) = \frac{1}{n_i} \sum_{j: y_j = i} I(f(x_j) \neq i),$$

where n_i is the number of points from z^n with $y = i$. We start by considering the almost obvious empirical optimization problem

$$\begin{array}{ll} \min_{f \in \hat{\mathcal{F}}_1} & \hat{e}_0(f) \\ \text{where} & \hat{\mathcal{F}}_1 = \{f : f \in \mathcal{F}, \hat{e}_1(f) \leq \alpha\}. \end{array} \tag{4}$$

Let $\hat{f}$ be a solution to (4) and let $e_0^* = \inf_{f \in \mathcal{F}_1} e_0(f)$ be the optimal error. In the Neyman–Pearson problem we are concerned with two estimation errors, $e_0(\hat{f}) - e_0^*$ and $e_1(\hat{f}) - \alpha$. We would like these to be small simultaneously. In addition, although we may encounter solutions where one of the two is negative we are only concerned with their size when they are positive.

Although we cannot guarantee with certainty that a solution $\hat{f}$ to this problem satisfies the constraint $e_1(\hat{f}) \le \alpha$, we may increase the probability of this event by changing the constraint in (4) to $\hat{e}_1(\hat{f}) \le \acute{\alpha}$ where $\acute{\alpha} < \alpha$. In doing so however we may restrict the size of $\hat{\mathcal{F}}_1$ which in turn may lower the probability that $e_0(\hat{f})$ is small. In fact with $\acute{\alpha} < \alpha$ it seems unlikely that we can control the estimation error $e_0(\hat{f}) - e_0^*$ in a distribution independent way. On the other hand we can control the estimation error in a rather straightforward manner if we allow $\acute{\alpha}$ to be slightly larger than α. That is, for a small price in the asymptotic value of the class 1 error we can obtain rates of convergence for both estimation errors simultaneously. We accomplish this by choosing $\acute{\alpha} = \alpha + \epsilon_1/2$ for some small $\epsilon_1 > 0$ and now let our candidate classifier $\hat{f}$ denote a solution of the slightly modified optimization problem

$$\begin{array}{ll} \min_{f \in \hat{\mathcal{F}}_1} & \hat{e}_0(f) \\ \text{where} & \hat{\mathcal{F}}_1 = \{f : f \in \mathcal{F}, \hat{e}_1(f) \le \alpha + \epsilon_1/2\}. \end{array} \tag{5}$$

and then derive bounds for

$$\mathcal{P}_S\Big(\big(e_0(\hat{f}) - e_0^* > \epsilon_0\big) \text{ or } \big(e_1(\hat{f}) > \alpha + \epsilon_1\big)\Big).$$

Although the Neyman–Pearson criterion ignores the class marginals the performance may not. This depends on how the n data samples are collected. The following fundamental lemma does not depend on the sample plan. That will become important later when we turn the consequences of this lemma into bounds on estimation error.

Lemma 1. *Consider two sets $\mathcal{X}$ and $\mathcal{Y}$. Let $e_0 : \mathcal{X} \to \Re$ and $e_1 : \mathcal{X} \to \Re$ denote two real univariate functions and let $\hat{e}_0 : \mathcal{X} \times \mathcal{Y} \to \Re$ and $\hat{e}_1 : \mathcal{X} \times \mathcal{Y} \to \Re$ denote two real bivariate functions.*

Let $\epsilon_0 > 0$, $\epsilon_1 > 0$, and $\alpha \ge 0$ be fixed and define

$$\mathcal{X}_1 = \{x \in \mathcal{X} : e_1(x) \le \alpha\}$$

$$\hat{\mathcal{X}}_1 = \{x \in \mathcal{X} : \hat{e}_1(x, y) \le \alpha + \epsilon_1/2\},$$

where we note that $\hat{\mathcal{X}}_1$ is a y-dependent subset of $\mathcal{X}$.

Let

$$x_* \in \arg\min_{x \in \mathcal{X}_1} e_0(x)$$

$$\hat{x}_*(y) \in \arg\min_{x \in \hat{\mathcal{X}}_1} \hat{e}_0(x, y)$$

where we write $\hat{x}_*(y)$ to emphasize that it is a function of y. Define the sets

$$\Theta_0 = \{y : e_0(\hat{x}_*(y)) - e_0(x_*) > \epsilon_0\}$$

$$\Theta_1 = \{y : e_1(\hat{x}_*(y)) > \alpha + \epsilon_1\}$$

$$\Omega_0 = \{y : \sup_{x \in \mathcal{X}_1} |e_0(x) - \hat{e}_0(x, y)| > \epsilon_0/2\}$$

$$\Omega_1 = \{y : \sup_{x \in \mathcal{X}_1} |e_1(x) - \hat{e}_1(x, y)| > \epsilon_1/2\}.$$

Then

$$\Theta_0 \cup \Theta_1 \ \subset \Omega_0 \cup \Omega_1.$$

Proof. Define the sets

$$C = \{y : \hat{e}_1(x_*, y) > \alpha + \epsilon_1/2\} \text{ and } \bar{C} = \{y : \hat{e}_1(x_*, y) \leq \alpha + \epsilon_1/2\}.$$

Then

$$\Theta_0 \cup \Theta_1 = (\Theta_0 \cap \bar{C}) \cup (\Theta_0 \cap C) \cup \Theta_1 \ \subset \ (\Theta_0 \cap \bar{C}) \cup C \cup \Theta_1 \ . \tag{6}$$

We consider the three terms on the right hand side separately. For the first term $\Theta_0 \cap \bar{C}$, $\bar{C}$ implies that $\hat{e}_1(x_*, y) \leq \alpha + \epsilon_1/2$ which implies that $x_* \in \hat{\mathcal{X}}_1$ and consequently $\hat{e}_0(x_*, y) \geq \hat{e}_0(\hat{x}_*(y), y)$. Therefore

$$\begin{aligned} e_0(\hat{x}_*(y)) - e_0(x_*) &= e_0(\hat{x}_*(y)) - \hat{e}_0(\hat{x}_*(y), y) + \hat{e}_0(\hat{x}_*(y), y) - e_0(x_*) \\ &\leq e_0(\hat{x}_*(y)) - \hat{e}_0(\hat{x}_*(y), y) + \hat{e}_0(x_*, y) - e_0(x_*) \\ &\leq 2 \sup_{x \in \mathcal{X}_1} |e_0(x) - \hat{e}_0(x, y)|. \end{aligned}$$

and so

$$\Theta_0 \cap \bar{C} \ \subset \Omega_0. \tag{7}$$

For the second term C,

$$\hat{e}_1(x_*, y) > \alpha + \epsilon_1/2$$

if and only if

$$\hat{e}_1(x_*, y) - e_1(x_*) > \alpha + \epsilon_1/2 - e_1(x_*),$$

but since $x_* \in \mathcal{X}_1$ implies that $e_1(x_*) \leq \alpha$ we obtain

$$\alpha + \epsilon_1/2 - e_1(x_*) \geq \alpha + \epsilon_1/2 - \alpha = \epsilon_1/2.$$

Consequently

$$C \ \subset \ \{y : \hat{e}_1(x_*, y) - e_1(x_*) > \epsilon_1/2\} \ \subset \Omega_1. \tag{8}$$

For the third term Θ_1,

$$e_1(\hat{x}_*(y)) > \alpha + \epsilon_1$$

if and only if

$$e_1(\hat{x}_*(y)) - \hat{e}_1(\hat{x}_*(y), y) > \alpha + \epsilon_1 - \hat{e}_1(\hat{x}_*(y), y),$$

but since $\hat{x}_*(y) \in \hat{\mathcal{X}}_1$ implies that $\hat{e}_1(\hat{x}_*(y), y) \leq \alpha + \epsilon_1/2$ we obtain

$$\alpha + \epsilon_1 - \hat{e}_1(\hat{x}_*(y), y) \geq \alpha + \epsilon_1 - (\alpha + \epsilon_1/2) = \epsilon_1/2.$$

Consequently,

$$\Theta_1 \ \subset \ \{y : e_1(\hat{x}_*(y)) - \hat{e}_1(\hat{x}_*(y), y) > \epsilon_1/2\} \ \subset \ \Omega_1. \tag{9}$$

The proof is completed by substituting (7),(8), and (9) into (6). ♦

We can now prove a corollary which will be our basic tool.

Corollary 1. *Let $\epsilon_0 > 0$, $\epsilon_1 > 0$, and $\alpha > 0$ be fixed and let f_* be a solution of the Neyman-Pearson problem (3). Consider any sample plan S for generating n-samples and given an n-sample let $\hat{f}$ be a solution to the modified empirical Neyman-Pearson problem (5). Then*

$$\mathcal{P}_S\Big(\big(e_0(\hat{f}) - e_0(f_*) > \epsilon_0\big) \text{ or } \big(e_1(\hat{f}) > \alpha + \epsilon_1\big)\Big) \le$$

$$\mathcal{P}_S\Big(\sup_{f\in\mathcal{F}} |e_0(f) - \hat{e}_0(f)| > \epsilon_0/2\Big) + \mathcal{P}_S\Big(\sup_{f\in\mathcal{F}} |e_1(f) - \hat{e}_1(f)| > \epsilon_1/2\Big)$$

Proof. Applying lemma 1 with $\mathcal{X} = \mathcal{F}$ and $\mathcal{Y} = Z^n$ and using a union bound on the two resulting sets finishes the proof. ⧫

We now get more specific about the sample design S. Theorem 1 provides bounds for estimation error for retrospective sampling and Theorem 2 for *i.i.d.* sampling.

Theorem 1. *Under the assumptions of Corollary 1, consider the retrospective sample plan S. That is we choose n_0 samples with $y = 0$ i.i.d. with respect to $\mathcal{P}_0$ and n_1 samples with $y = 1$ i.i.d. with respect to $\mathcal{P}_1$ where $n = n_0 + n_1$. Suppose that $n_i \ge \frac{2}{\epsilon^2}$, $i = 0, 1$. Then*

$$\mathcal{P}_S\Big(\big(e_0(\hat{f}) - e_0(f^*) > \epsilon_0\big) \text{ or } \big(e_1(\hat{f}) > \alpha + \epsilon_1\big)\Big) \le$$

$$8n_0^{V(\mathcal{F})} e^{-n_0\epsilon_0^2/128} + 8n_1^{V(\mathcal{F})} e^{-n_1\epsilon_1^2/128}.$$

where $V(\mathcal{F})$ is the Vapnik-Chervonenkis dimension of $\mathcal{F}$.

Proof. With the retrospective sample plan $\mathcal{P}_S = \mathcal{P}_0^{n_0}\mathcal{P}_1^{n_1}$, so the right hand side of Corollary 1 satisfies

$$\mathcal{P}_S\Big(\sup_{f\in\mathcal{F}} |e_i(f) - \hat{e}_i(f)| > \rho\Big) = \mathcal{P}_i^{n_i}\Big(\sup_{f\in\mathcal{F}} |e_i(f) - \hat{e}_i(f)| > \rho\Big).$$

The proof then follows directly from the empirical process convergence theorem of (Vapnik & Chervonenkis, 1974)

$$\mathcal{P}^m(\sup_{f\in\mathcal{F}} |e(f) - \hat{e}(f)| > \epsilon) \le 8m^{V(\mathcal{F})} e^{-m\epsilon^2/32} \tag{10}$$

when $m \ge \frac{2}{\epsilon^2}$ (see also (Devroye, Györfi, & Lugosi, 1996; Vapnik, 1998)). ⧫

We now address *i.i.d* sampling. Then n_i are random variables, but we use their concentration about their means to obtain a similar result. We utilize the following lemma.

Lemma 2. *Let $Z = (X, Y)$ be a probability space where $Y = \{0, 1\}$, with probability measure $\mathcal{P}$, conditional probability measure $\mathcal{P}_0 = \mathcal{P}(\cdot|y = 0)$ and marginal $p = P(y = 0)$. Let $z = (x, y)$ denote the corresponding random variable. Let $\rho > 0$. Let $z^n = \{z_i, i = 1, .., n\} \in Z^n$ denote an* i.i.d. *n-sample with $n \geq \frac{10\sqrt{5}}{p^2\rho^2}$. Let $\mathcal{F}$ be a class of functions $f : X \to Y$ and let*

$$\begin{aligned} e_0(f) &= \mathcal{P}_0\big(f(x) \neq 0\big) \\ \hat{e}_0(f) &= \frac{1}{n_0} \sum_{j:y_j=0} I(f(x_j) \neq 0) \end{aligned}$$

where n_0 denotes the number of samples with $y = 0$. Then

$$\mathcal{P}_S\Big(\sup_{f\in\mathcal{F}} |e_0(f) - \hat{e}_0(f)| > \rho\Big) \leq 10(2n)^{V(\mathcal{F})} e^{-\frac{np^2\rho^2}{160\sqrt{5}}}.$$

Proof. Define the sets

$$\begin{aligned} A &= \left\{ z^n : \sup_{f\in\mathcal{F}} |e_0(f) - \hat{e}_0(f)| > \rho \right\} \\ B &= \left\{ z^n : \Big|\frac{n_0}{n} - p\Big| > p\gamma \right\} \end{aligned}$$

for some $1 \geq \gamma \geq 0$ to be determined. Then

$$\begin{aligned} \mathcal{P}_S(A) \quad &= \mathcal{P}_S(A|B)\mathcal{P}_S(B) + \mathcal{P}_S(A|\bar{B})\mathcal{P}_S(\bar{B}) \\ &\leq \mathcal{P}_S(B) + \mathcal{P}_S(A|\bar{B}) \end{aligned} \tag{11}$$

Since n_0 is a binomially distributed random variable with parameters n and p, we can use the bound of (Chernoff, 1952) (see also Chapter 8 in (Devroye *et al.*, 1996)) to obtain

$$\mathcal{P}_S(B) = \mathcal{P}_S\left(\Big|\frac{n_0}{n} - p\Big| > p\gamma\right) \leq 2e^{-2np^2\gamma^2} \tag{12}$$

To bound the term $\mathcal{P}_S(A|\bar{B})$ let

$$M = \left\{ m : \Big|\frac{m}{n} - p\Big| \leq p\gamma \right\}$$

so that

$$np(1-\gamma) \leq m \leq np(1+\gamma), \quad m \in M. \tag{13}$$

We suppose for the moment that

$$np(1-\gamma) \geq \frac{2}{\rho^2}. \tag{14}$$

Then by the VC theorem (10) we can bound

$$\begin{aligned} \mathcal{P}_S(A|\bar{B}) \quad &= \frac{\sum_{m\in M} \mathcal{P}_S\Big(\sup_{f\in\mathcal{F}}|e_0(f)-\hat{e}_0(f)|>\rho \Big| n_0=m\Big) P(n_0=m)}{\sum_{m\in M} P(n_0=m)} \\ &\leq \max_{m\in M} \mathcal{P}_0^{n_0}\Big(\sup_{f\in\mathcal{F}} |e_0(f) - \hat{e}_0(f)| > \rho \Big| n_0 = m\Big) \\ &\leq \max_{m\in M} \left(8m^{V(\mathcal{F})} e^{-m\rho^2/32}\right). \end{aligned} \tag{15}$$

Let $m_l = np(1-\gamma)$ be a lower bound on the smallest member of M and $m_u = np(1+\gamma)$ be an upper bound on the largest member of M so that

$$\max_{m \in M} \left(8m^{V(\mathcal{F})}e^{-m\rho^2/32}\right) \leq 8m_u^{V(\mathcal{F})}e^{-m_l\rho^2/32}.$$

Substituting this into (15) and combining with (12) and (11) gives

$$\mathcal{P}_S(A) \leq 2e^{-2np^2\gamma^2} + 8\big(np(1+\gamma)\big)^{V(\mathcal{F})}e^{-np(1-\gamma)\rho^2/32}. \tag{16}$$

A larger γ makes the first term smaller and the second term larger. We choose γ to approximately equate the two exponentials by solving

$$p^2\gamma^2 = p(1-\gamma).$$

We use the solution $\gamma = \frac{-1+\sqrt{1+4p}}{2p}$ and simplify the resulting exponentials by bounding

$$p^2\gamma^2 = p(1-\gamma) = \frac{1}{2}(1 + 2p - \sqrt{1+4p})$$

from below. To do this we denote $f(p) = 1 + 2p - \sqrt{1+4p}$ and note that $f(0) = \dot{f}(0) = 0$ and $\ddot{f}(p) = 4(1+4p)^{-\frac{3}{2}} \geq \frac{4}{5\sqrt{5}}$ for $0 \leq p \leq 1$. Therefore by Taylor's theorem

$$f(p) \geq \frac{2}{5\sqrt{5}}p^2$$

so that with this choice of γ

$$p^2\gamma^2 = p(1-\gamma) \geq \frac{p^2}{5\sqrt{5}}.$$

Consequently, from the assumption $n \geq \frac{10\sqrt{5}}{p^2\rho^2}$ of the lemma, (14) is satisfied. In addition we can also bound

$$p(1+\gamma) \leq 2.$$

Since $\frac{\rho^2}{32} \leq 2$ we can now simplify the inequality (16) to obtain

$$\mathcal{P}_S(A) \leq \big(2 + 8(2n)^{V(\mathcal{F})}\big)e^{-\frac{np^2\rho^2}{160\sqrt{5}}} \leq 10(2n)^{V(\mathcal{F})}e^{-\frac{np^2\rho^2}{160\sqrt{5}}}$$

and the proof is finished.

♦

We can now prove

Theorem 2. *Under the assumptions of Corollary* 1*, consider taking n i.i.d. samples with $n \geq \frac{10\sqrt{5}}{p_i^2\epsilon_i^2}$, $i = 0, 1$ where $p_i = P(y = i)$ are the class marginals. Then*

$$\mathcal{P}_S\Big(\big(e_0(\hat{f}) - e_0(f^*) > \epsilon_0\big) \text{ or } \big(e_1(\hat{f}) > \alpha + \epsilon_1\big)\Big) \leq$$

$$10(2n)^{V(\mathcal{F})}\big(e^{-\frac{np_0^2\epsilon_0^2}{640\sqrt{5}}} + e^{-\frac{np_1^2\epsilon_1^2}{640\sqrt{5}}}\big).$$

Proof. We apply Corollary 1 and apply Lemma 2, with respect to both $y = 0$ and $y = 1$, to the consequential terms and the proof is finished. ⧫

3 The min–max Problem

In this section we introduce and analyze a min–max criterion that forms the foundation for a learning paradigm we call, following (Huber, 1972), *robust machine learning*. We start with a very general setting that includes the design of all types of predictors (not just classifiers). In this setting the design criterion e is defined in terms of an unknown parameter q and we wish to design a predictor that is robust to its value. The min–max approach consists in designing the predictor by solving

$$\min_{f \in \mathcal{F}} \max_{q \in Q} e(f, q). \tag{17}$$

Let us suppose that we can solve this min–max problem with a solution f^* and optimal value e^*. Then for whatever value q that is chosen, by nature or the opponent, we are guaranteed that

$$e(f^*, q) \leq e^*.$$

Therefore, although such a guarantee may not be nearly as good as one could get if one knew q, it is protection against possible catastrophic losses obtainable when designing a classifier by minimizing $e(f, q)$ with the incorrect value of q. This standard game theoretic argument changes when we discuss learning from samples. Indeed, then we need to specify both the sample design and the empirical loss function $\hat{e}(f, q)$ as a function of the n-sample for every $q \in Q$. Although there appears to be no general way of doing this, we discuss three situations where we know how to proceed and analyze the third in some detail.

In the first case let us consider where q represents some information about the measure $\mathcal{P}$. To emphasize this dependence we write $\mathcal{P}_q$. For exposition purposes let us consider classification. Suppose now the sample plan samples *i.i.d.* from $\mathcal{P}_{q_{sample}}$ but the generalization error is computed on future data where the q has drifted to some q_{future}. We do not know q_{sample} and q_{future} but know that the drift is less than ε in some metric d on Q. That is, $d(q_{sample}, q_{future}) \leq \varepsilon$. Given an n sample z^n from $\mathcal{P}_{q_{sample}}$ we would like to compute a reasonable empirical approximation $\hat{e}(f, q_{future})$. Although we do not know anything about this problem in general we do know a way to proceed when q is the x mean and d is determined by a norm $|\cdot|$. Then to compute $\hat{e}(f, q)$ for all q with $d(q_{sample}, q) \leq \varepsilon$ we could compute the usual empirical error with respect to the shifted n-sample $\big((x_1 + \delta q, y_1).(x_2 + \delta q, y_2, .., (x_n + \delta q, y_n)\big)$ for all δq with $|\delta q| \leq \varepsilon$ and call this $\hat{e}(f, q_{sample} + \delta q)$. We could then solve

$$\min_{f \in \mathcal{F}} \max_{|\delta q| \leq \varepsilon} \hat{e}(f, q_{sample} + \delta q) \tag{18}$$

and try to prove guarantees about the generalization error compared to the optimal $e^*(q_{future})$ at q_{future}.

In the second case, the sample plan S is *i.i.d.* from some unknown measure $\mathcal{P}$ and the model error is determined by integration of a loss function with respect to the same $\mathcal{P}$ but the loss function is specified in terms of a parameter $q \in Q$. This format is useful when the customer wants the practicioner to build a model but cannot seem to specify a loss function. We could call this a technique for building models that are robust to customer capriciousness.

The third and most common case is when q represents the class marginals in classification and the sample plan is retrospective. We discuss this situation in more detail in the next section. First we have more to say about the general situation.

We now consider where we wish to solve the min–max problem (17). We collect n samples through some sample plan S, determine an empirical version $\hat{e}$ of the loss e in some way and approximately solve this min–max problem by solving its empirical version

$$\min_{f \in \mathcal{F}} \max_{q \in Q} \hat{e}(f, q). \tag{19}$$

We would like guarantees on how close the approximate solution is to a true solution of (17).

The main tool we use is the following fundamental lemma for min–max. It is important to observe that although we utilize this lemma in learning, it is a much more general fact concerning how well a solution to a nearby min–max problem solves a min–max problem.

Lemma 3. *Consider two sets $\mathcal{X}$ and $\mathcal{Y}$ and let $e : \mathcal{X} \times \mathcal{Y} \to \Re$ and $\hat{e} : \mathcal{X} \times \mathcal{Y} \to \Re$ denote two real bivariate functions. Let*

$$e^* = \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} e(x, y) \tag{20}$$

and

$$e_* = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} e(x, y) \tag{21}$$

denote the min–max and max–min values respectively, and let $\hat{x}$ be a solution to the approximate min–max problem

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \hat{e}(x, y). \tag{22}$$

Then for any $y \in \mathcal{Y}$

$$e(\hat{x}, y) - e^* \leq 2 \sup_{x,y} |e(x, y) - \hat{e}(x, y)| \tag{23}$$

and there exists a y^* such that

$$e_* \ \leq \ e(\hat{x}, y^*) \ \leq \ e^* + 2 \sup_{x,y} |e(x, y) - \hat{e}(x, y)|. \tag{24}$$

Proof. Let $\hat{e}^*$ denote the value of (22). Decompose

$$e(\hat{x}, y) - e^* = e(\hat{x}, y) - \hat{e}(\hat{x}, y) + \hat{e}(\hat{x}, y) - e^* \tag{25}$$

and bound the last two terms $\hat{e}(\hat{x}, y) - e^*$. Since

$$\hat{e}(\hat{x}, y) \leq \max_{y \in \mathcal{Y}} \hat{e}(\hat{x}, y) = \hat{e}^*,$$

$$\hat{e}(\hat{x}, y) - e^* \leq \hat{e}^* - e^*$$

we only need to control $\hat{e}^* - e^*$. Now

$$\hat{e}^* - e^* = \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \hat{e}(x, y) - \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} e(x, y) \leq \max_{y \in \mathcal{Y}} \hat{e}(x^*, y) - \max_{y \in \mathcal{Y}} e(x^*, y).$$

Let $\bar{y}$ denote a solution to

$$\max_{y \in \mathcal{Y}} \hat{e}(x^*, y) = \hat{e}(x^*, \bar{y}).$$

Then

$$\max_{y \in \mathcal{Y}} \hat{e}(x^*, y) - \max_{y \in \mathcal{Y}} e(x^*, y) \leq \hat{e}(x^*, \bar{y}) - e(x^*, \bar{y})$$

so that substitution into (25) gives

$$\begin{aligned}
& e(\hat{x}, y) - e^* \leq e(\hat{x}, y) - \hat{e}(\hat{x}, y) + \hat{e}(x^*, \bar{y}) - e(x^*, \bar{y}) \\
& \leq \sup_{x,y} \left(e(x, y) - \hat{e}(x, y)\right) + \sup_{x,y} \left(\hat{e}(x, y) - e(x, y)\right) \\
& = \sup_{x,y} \left(e(x, y) - \hat{e}(x, y)\right) - \inf_{x,y} \left(e(x, y) - \hat{e}(x, y)\right) \\
& \leq 2 \sup_{x,y} |e(x, y) - \hat{e}(x, y)|
\end{aligned}$$

and the first statement is proved. To prove the second statement we define y^* to be a solution of the max–min problem (21) and observe that

$$e_* = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} e(x, y) = \min_{x \in \mathcal{X}} e(x, y^*) \leq e(\hat{x}, y^*).$$

⧫

We are now in a position to prove a corollary which will be our main tool in providing performance guarantees for min–max.

Corollary 2. *Let e^* denote the value of the min–max problem (17), $e_* = \max_{q \in Q} \min_{f \in \mathcal{F}} e(f, q)$ the value of the max–min problem, and let a sample plan S be specified. For a given n-sample let $\hat{f}$ denote a solution of the empirical min–max problem (19). Then for any q and any $\epsilon > 0$*

$$\mathcal{P}_S\Big(e(\hat{f}, q) - e^* > \epsilon\Big) \leq \mathcal{P}_S\Big(\sup_{f \in \mathcal{F}, q \in Q} |e(f, q) - \hat{e}(f, q)| > \epsilon/2\Big)$$

and there exists a q^* such that

$$\mathcal{P}_S\Big(e_* \leq e(\hat{f}, q^*) \leq e^* + \epsilon\Big) \geq \mathcal{P}_S\Big(\sup_{f \in \mathcal{F}, q \in Q} |e(f, q) - \hat{e}(f, q)| \leq \epsilon/2\Big).$$

Proof. The proof follows directly from Lemma 3 with the choices $\mathcal{X} = \mathcal{F}$ and $\mathcal{Y} = Q$. ⧫

3.1 Min–max Over Class Marginals

To utilize Corollary 2 for performance guarantees more structure needs to be specified. Here we show how to treat the most common min–max problem. In particular, we consider classification with $M = |Y|$ classes, where the sample design is retrospective and we have no information about the Y marginals. Let

$$Q = \{(q_1, q_2, ..., q_M) : q_i \geq 0, \sum_{i=1}^{M} q_i = 1\}.$$

If the Y marginals were $q \in Q$, then the generalization error

$$e(f, q) = \sum_{i=1}^{M} e_i(f) q_i,$$

where we recall the conditional class error functions

$$e_i(f) = \mathcal{P}_i\big(f(x) \neq i\big),\ i = 1, .., M.$$

For the retrospective sample with n_i samples in class $i \in Y$ we let the empirical error be defined in the natural way

$$\hat{e}(f, q) = \sum_{i=1}^{M} \hat{e}_i(f) q_i$$

where we use the empirical conditional class errors

$$\hat{e}_i(f) = \frac{1}{n_i} \sum_{j: y_j = i} I(f(x_j) \neq i),\ i = 1, .., M.$$

Although this definition seems sensible, it depends upon the class marginals $q \in Q$. Therefore any classifier built by minimizing this empirical error will depend on our choice of q. Since we evaluate the performance of a classifier with a specific unknown q chosen by nature, we would like build a classifier that will not suffer too much from an adversarial nature. One way to accomplish this is through the min–max technique. Due to the structure of these loss functions, the min–max (17) and empirical min–max (19) problems become

$$\min_{f \in \mathcal{F}} \max_{i \in Y} e_i(f) \tag{26}$$

and

$$\min_{f \in \mathcal{F}} \max_{i \in Y} \hat{e}_i(f). \tag{27}$$

We now state a performance bound for this min–max problem.

Theorem 3. *Consider a retrospective sample plan S of size n with n_i samples drawn i.i.d. from $\mathcal{P}_i$. Let e^* denote the value of the min–max problem (26), $e_* = \max_{q \in Q} \min_{f \in \mathcal{F}} \sum_{i=1}^M e_i(f) q_i$ the value of the max–min problem, and let $\hat{f}$ denote a solution to the empirical min–max problem (27). Let $\mathcal{F}_i, i = 1, .., M$ denote the space $\mathcal{F}$ mapped in the natural way to binary classifiers i or not i. Then for any $q \in Q$,*

$$\mathcal{P}_S\Big(\sum_{i=1}^M e_i(\hat{f}) q_i - e^* > \epsilon\Big) \leq M \max_{i \in Y} n_i^{V(\mathcal{F}_i)} e^{-n_i \epsilon^2/128} \tag{28}$$

and there exists a $q^* \in Q$ such that

$$\mathcal{P}_S\Big(e_* \leq \sum_{i=1}^M e_i(\hat{f}) q_i^* \leq e^* + \epsilon\Big) \geq 1 - M \max_{i \in Y} n_i^{V(\mathcal{F}_i)} e^{-n_i \epsilon^2/128}.$$

Proof. The proof follows from Corollary 2 and

$$\begin{aligned}\mathcal{P}_S\Big(\sup_{f \in \mathcal{F}} \max_{i \in Y} |e_i(f) - \hat{e}_i(f)| > \epsilon/2\Big) &\leq M \max_{i \in Y} \mathcal{P}_S\Big(\sup_{f \in \mathcal{F}} |e_i(f) - \hat{e}_i(f)| > \epsilon/2\Big) \\ &= M \max_{i \in Y} \mathcal{P}_i^{n_i}\Big(\sup_{f \in \mathcal{F}_i} |e_i(f) - \hat{e}_i(f)| > \epsilon/2\Big) \\ &\leq M \max_{i \in Y} n_i^{V(\mathcal{F}_i)} e^{-n_i \epsilon^2/128},\end{aligned}$$

where the last line follows from the VC theorem (10). ♦

References

Chernoff, H. (1952). A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, *23*, 493–507.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, New York, NY.

Huber, P. J. (1972). Robust statistics: a review. *The Annals of Mathematical Statistics*, *43*, 1041–1067.

Van Trees, H. L. (1968). *Detection, Estimation and Modulation Theory, Part I.* John Wiley and Sons, New York.

Vapnik, V. N. (1998). *Statistical Learning Theory.* John Wiley and Sons, Inc., New York.

Vapnik, V. N., & Chervonenkis, A. (1974). *Theory of Pattern Recognition.* Nauka, Moscow. (in Russian).